REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
07-03-2012	Final Technical	06/01/20009--11/30/2011; Mar 2012

4. TITLE AND SUBTITLE

Information Fusion: Inference from Graphs and Feature Matrices

5a. CONTRACT NUMBER

5b. GRANT NUMBER

FA9550-09-1-0399

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Priebe, Carey E.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218-2686

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Science and Research
875 Randolph Street
Suite 325, Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-OSR-VA-TR-2012-0546

12. DISTRIBUTION / AVAILABILITY STATEMENT

A

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Multiple information sources -- feature matrices for each entity as well as graph structure between entities -- need to be fused for inferential exploitation. Methodologies for joint inference in such disparate information settings have been developed. In particular, the methodologies introduced in the manuscripts "Dimensionality Reduction on the Cartesian Product of Embeddings of Multiple Dissimilarity Matrices" (Journal of Classification, 2010) and "Manifold Matching: Joint Optimization of Fidelity and Commensurability" (Brazilian Journal of Probability and Statistics, accepted for publication, 2012) are of fundamental and far-reaching importance. Indeed, this work forms the theoretical and methodological core for the major continuing thrust of the PI's research program.
(These manuscripts are available at
http://www.cis.jhu.edu/~parky/CEP-Publications/journal.html)

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
unclassified	unclassified	unclassified			

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Manifold Matching: Joint Optimization of Fidelity and Commensurability

Carey E. Priebe*, Johns Hopkins University
David J. Marchette, Naval Surface Warfare Center
Zhiliang Ma, Johns Hopkins University
Sancar Adali, Johns Hopkins University

November 12, 2011

Abstract

Fusion and inference from multiple and massive disparate data sources – the requirement for our most challenging data analysis problems and the goal of our most ambitious statistical pattern recognition methodologies – has many and varied aspects which are currently the target of intense research and development. One aspect of the overall challenge is manifold matching – identifying embeddings of multiple disparate data spaces into the same low-dimensional space where joint inference can be pursued. We investigate this manifold matching task from the perspective of jointly optimizing the fidelity of the embeddings and their commensurability with one another, with a specific statistical inference exploitation task in mind. Our results demonstrate when and why our joint optimization methodology is superior to either version of separate optimization. The methodology is illustrated with simulations and an application in document matching.

1 Introduction

1.1 Motivation

Let $(\Xi, \mathcal{F}, \mathcal{P})$ be a probability space, i.e., Ξ is a sample space, $\mathcal{F}$ is a sigma-field, and $\mathcal{P}$ is a probability measure. Consider K measurable spaces $\Xi_1, \cdots, \Xi_K$ and measurable maps $\pi_k : \Xi \to \Xi_k$. Each π_k induces a probability measure $\mathcal{P}_k$ on Ξ_k. We wish to identify a measurable metric space $\mathcal{X}$ (with distance function d) and measurable maps $\rho_k : \Xi_k \to \mathcal{X}$, inducing probability measures $\widetilde{\mathcal{P}}_k$ on $\mathcal{X}$, so that for $[x_1, \cdots, x_K]' \in \Xi_1 \times \cdots \times \Xi_K$ we may evaluate distances $d(\rho_{k_1}(x_{k_1}), \rho_{k_2}(x_{k_2}))$ in $\mathcal{X}$. See Figure 1.

Given $\xi_1, \xi_2 \overset{iid}{\sim} \mathcal{P}$ in Ξ, we may reasonably hope that the random variable $d(\rho_{k_1} \circ \pi_{k_1}(\xi_1), \rho_{k_2} \circ \pi_{k_2}(\xi_1))$ is stochastically smaller than the random variable $d(\rho_{k_1} \circ \pi_{k_1}(\xi_1), \rho_{k_2} \circ \pi_{k_2}(\xi_2))$. That is,

*Corresponding Author: Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218-2682 ; cep@jhu.edu . This work is partially supported by National Security Science and Engineering Faculty Fellowship (NSSEFF), Air Force Office of Scientific Research (AFOSR), Office of Naval Research (ONR), Johns Hopkins University Human Language Technology Center of Excellence (JHU HLT COE), and the American Society for Engineering Education (ASEE) Sabbatical Leave Program.

matched measurements $\pi_{k_1}(\xi_1), \pi_{k_2}(\xi_1)$ representing a single point ξ_1 in Ξ are mapped closer to each other than are unmatched measurements $\pi_{k_1}(\xi_1), \pi_{k_2}(\xi_2)$ representing two different points in Ξ. This property allows inference to proceed in the common representation space $\mathcal{X}$.

However, we do not observe $\xi \in \Xi$; we also do not observe the $x_k = \pi_k(\xi) \in \Xi_k$ directly, nor do we have knowledge of the maps π_k. But suppose we have access to functions $\delta_k : \Xi_k \times \Xi_k \to \mathbb{R}_+ = [0, \infty)$ such that $\delta_k(\pi_k(\xi_1), \pi_k(\xi_2))$ represents the "dissimilarity" of outcomes ξ_1 and ξ_2 under map π_k. We propose to use sample dissimilarities for matched data in the disparate spaces Ξ_k to simultaneously learn maps ρ_k which allow for a powerful test of matchedness in the common representation space $\mathcal{X}$.

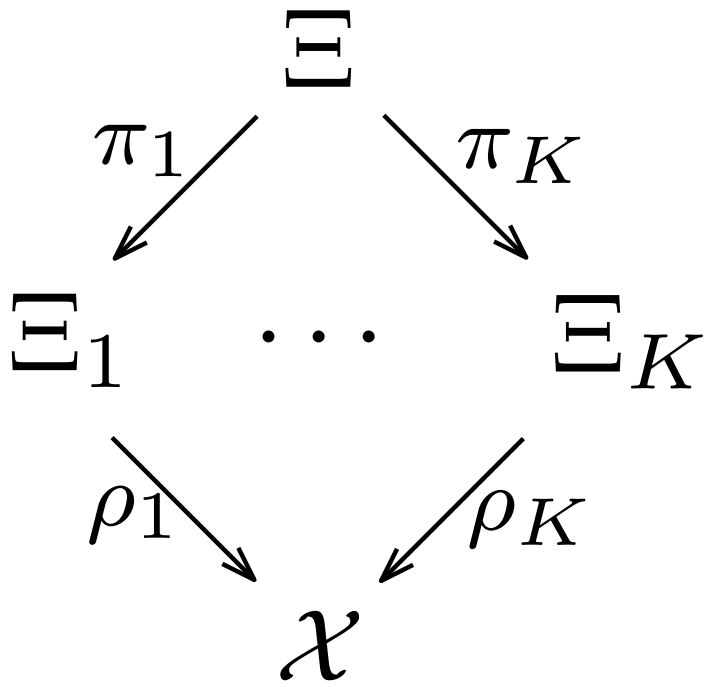

Figure 1: Maps π_k induce disparate data spaces Ξ_k from "object space" Ξ. Manifold matching involves using matched data $\{\boldsymbol{x}_{ik}\}$ to simultaneously learn maps $\rho_1, \ldots, \rho_K$ from disparate spaces $\Xi_1, \ldots, \Xi_K$ to a common "representation space" $\mathcal{X}$, for subsequent inference.

1.2 Problem Formulation

Consider n objects each measured under K different conditions,

$$\boldsymbol{x}_{i1} \sim \cdots \sim \boldsymbol{x}_{ik} \sim \cdots \sim \boldsymbol{x}_{iK}, \ i = 1, \ldots, n,$$

where $\boldsymbol{x}_{i1} \sim \cdots \sim \boldsymbol{x}_{ik} \sim \cdots \sim \boldsymbol{x}_{iK}$ denotes K matched measurements $\pi_1(\xi_i), \cdots, \pi_K(\xi_i)$ representing a single object $\xi_i \in \Xi$, where Ξ denotes the "object space". The assumption of K *different* conditions implies that $\boldsymbol{x}_{ik} \in \Xi_k$ where the spaces $\Xi_1, \cdots, \Xi_K$ cannot be assumed to be similar. We are given K new measurements $\{\boldsymbol{y}_k\}_{k=1}^K$, $\boldsymbol{y}_k \in \Xi_k$. The question under consideration is: Does the collection $\{\boldsymbol{y}_k\}_{k=1}^K$ also correspond to matched measurements representing a single object measured under the K conditions?

We use the Ξ notation to remind the reader that the spaces Ξ_k cannot be assumed to be standard finite-dimensional Euclidean spaces. We do assume that each space Ξ_k comes with a within-condition dissimilarity δ_k – a hollow, symmetric function from $\Xi_k \times \Xi_k$ to $\mathbb{R}_+$ – through which the

matched data $\{\boldsymbol{x}_{ik}\}$ yields $n \times n$ dissimilarity matrices Δ_k, $k = 1, \cdots, K$. For new measurements $\{\boldsymbol{y}_k\}_{k=1}^K$ we have available for each k the within-condition dissimilarities $\delta_k(\boldsymbol{y}_k, \boldsymbol{x}_{ik})$, $i = 1, \ldots, n$.

Remark 1: The $\boldsymbol{x}_{ik}$ and $\boldsymbol{y}_k$ are introduced mainly for symbolic purposes; the corresponding data may not be available or may be too complex to use directly, and we proceed from the dissimilarities.

The specific statistical inference exploitation task we consider throughout most of this article is hypothesis testing. Our goal, simplified for the case $K = 2$, is to determine whether $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are a match. That is,

$$H_0 : \boldsymbol{y}_1 \sim \boldsymbol{y}_2 \quad \text{versus} \quad H_A : \boldsymbol{y}_1 \nsim \boldsymbol{y}_2,$$

or equivalently,

$$H_0 : \boldsymbol{y}_1 = \pi_1(\xi), \boldsymbol{y}_2 = \pi_2(\xi) \quad \text{versus} \quad H_A : \boldsymbol{y}_1 = \pi_1(\xi), \boldsymbol{y}_2 = \pi_2(\xi') \quad \text{for} \quad \xi \neq \xi' \in \Xi.$$

(We control the probability of missing a true match.)

1.3 Manifold Matching

We define *manifold matching* as simultaneous manifold learning and manifold alignment – identifying embeddings of multiple disparate data sources into the same low-dimensional space where joint inference can be pursued. Figure 1 depicts our framework. Conditional distributions are induced by maps π_k from "object space" Ξ. Our assumption is that the conditional spaces Ξ_k are *not* commensurate. For example, if the elements of Ξ are individual people, then a photograph in image space Ξ_1 and a biographical sketch in text document space Ξ_2 are not to be directly compared. Indeed, our fundamental premise defining *disparate* data sources is that the various Ξ_k cannot profitably be treated as replicates of the same kind of space. Rather, the various spaces are different not just in degree but in kind. Each dissimilarity δ_k has been tailored for application to Ξ_k, and it is inappropriate to apply δ_k on $\Xi_k \times \Xi_{k'}$ for $k' \neq k$. This distinguishes our *data fusion* from conventional multivariate analysis.

In Figure 1, matched points $\{\boldsymbol{x}_{ik}\}$ are used to simultaneously learn appropriate maps ρ_k taking the disparate data from the various Ξ_k into a common representation space $\mathcal{X}$. These maps are then applied to $\{\boldsymbol{y}_k\}_{k=1}^K$ yielding $\widetilde{\boldsymbol{y}}_k = \rho_k(\boldsymbol{y}_k)$, whence (for $K = 2$) we use $T = d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$ as our test statistic and reject for T "large".

Remark 2: Our convention is to use the " $\widetilde{\cdot}$ " notation for points in the target space $\mathcal{X}$, contrasted with no tilde for points in the original Ξ_k spaces.

Remark 3: We will throughout consider the special case of $\mathcal{X} = \mathbb{R}^m$ for some pre-specified target dimension m. The fundamentally important and challenging task of choosing the target dimension – *model selection* – will be considered only as a confounding issue in this paper; m is a nuisance parameter which must be selected but whose selection is beyond the scope of this manuscript.

1.4 What are these "conditions" and what does "matched" mean?

As suggested above, one example of "conditions" involves photographs $\{\boldsymbol{x}_{i1}\}$ and biographical sketches $\{\boldsymbol{x}_{i2}\}$, with "matched" $\boldsymbol{x}_{i1} \sim \boldsymbol{x}_{i2}$ meaning that the photograph $\boldsymbol{x}_{i1}$ and the biographical sketch $\boldsymbol{x}_{i2}$ are of the same person.

Other illustrative examples include: a general image & caption scenario, with "matched" meaning that they go together; multiple languages for text documents, with "matched" meaning on the same topic; multiple modalities for photographs (e.g., indoor lighting vs outdoor lighting, two cameras of different quality, or passport photos and airport surveillance photos),

with "matched" meaning of the same person; Wikipedia text document and Wikipedia hyperlink structure, with "matched" meaning of the same document. More generally, our framework may be applicable to any scenario in which multiple dissimilarity measures are applied to the objects at hand.

Fundamentally, "matched" means whatever the training data say it means. We know it when we see it – or, perhaps more accurately, we know *un*matched when we see it; see Figure 2. Consider, for instance, an example of multiple languages for text documents, with "matched" meaning on the same topic. Given English and French Wikipedia documents with the matching provided by Wikipedia itself, "matched" means "on the same topic." But of course the Wikipedia documents are not direct translations of one another, and documents in different languages on the same topic may have significant conceptual differences due to cultural differences, etc.

Figure 2: An example of "not matched" for multi-lingual text documents. The English is clear enough to lorry drivers — but the Welsh reads "I am not in the office at the moment. Send any work to be translated." (See `http://news.bbc.co.uk/2/hi/uk_news/wales/7702913.stm`; permission obtained from `http://www.golwg360.com/Hafan/default.aspx`.)

1.5 Dirichlet Setting

While the matched training data ultimately determine what "matched" means, in order to provide a clear mathematical characterization of matchedness we consider an illustrative Dirichlet setting. This setting is clearly overly simplified, but it invokes some aspects of the foregoing example of multiple languages for text documents.

Let $S^p = \{\boldsymbol{x} \in \mathbb{R}_+^{p+1} : \sum_{\ell=1}^{p+1} x_\ell = 1\}$ be the standard p-simplex. We consider here the case $\Xi_1 = S^p$ and $\Xi_2 = S^p$ – the two spaces are, in fact, commensurate in this case, for illustration. Let $\boldsymbol{\gamma}_i \overset{iid}{\sim} Dirichlet(\mathbf{1})$ represent n "objects" or "topics". Let $X_{ik} \overset{iid}{\sim} Dirichlet(r\boldsymbol{\gamma}_i + \mathbf{1})$ represent document i in language k. (Since the X_{ik} take their value in S^p, we can think of them as modelling (normalized) word count histograms with $p+1$ distinct words. $\Xi_1 = \Xi_2 = S^p$ suggests a simplified 1-1 word correspondence model. A permutation σ indicating that the 1-1 word correspondence is unknown may be applied to the dimensions of one space with no alteration to our illustration.) In this case, r controls what it means to be matched – e.g., document

translation quality analogy. If r is large (highly accurate translations), then matched documents X_{i1} and X_{i2} will be probabilistically more similar than X_{i1} and $X_{i'2}$ for $i \neq i'$; if r is small (rough translations), then "matched" doesn't mean much. Indeed, the limiting case of $r \to \infty$ (point masses) yields "matched" means "identical" while $r = 0$ (recall that $Dirichlet(\mathbf{1})$ is uniform on the simplex) yields "matched" means "no relationship". Figure 3, with $p = 2$, provides an illustration wherein matched means quite a lot. A real data version of this setting with multiple documents per topic is depicted in Figure 4, where three Linguistic Data Consortium (LDC) Enron email message topic classes are projected into the simplex S^2 via Fisher's Linear Discriminant composed with Latent Semantic Analysis (FLD∘LSA) (see, e.g., [1, 2, 3]).

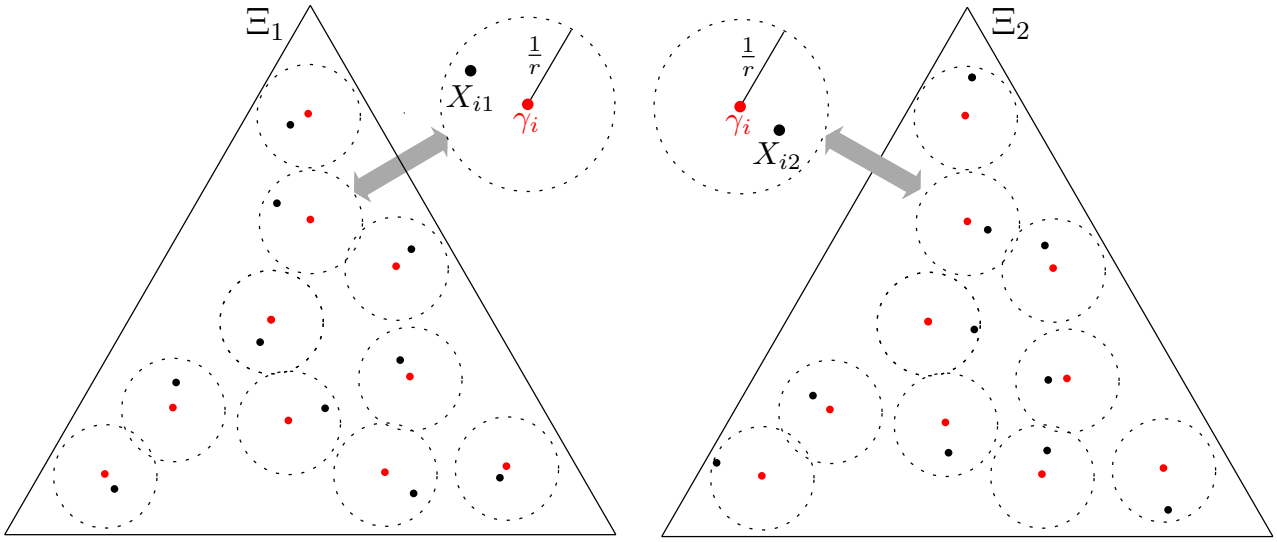

Figure 3: Illustrative Dirichlet setting wherein $X_{ik} \overset{iid}{\sim} Dirichlet(r\boldsymbol{\gamma}_i + \mathbf{1})$ represent documents $i = 1, \ldots, n = 10$ in languages $k = 1, \ldots, K = 2$ in the standard 2-simplex S^2. The parameter r controls the meaning of matchedness – the similarity of matched documents X_{i1} and X_{i2} compared to unmatched documents X_{i1} and $X_{i'2}$ for $i \neq i'$.

1.6 Related Work

The 2006 David Hand polemic [4] argued persuasively that a fundamental issue in statistical inference research and development – perhaps *the* fundamental issue – is robustness in the face of test data drawn from a distribution *not* the same as the distribution from which the training data are drawn. The disparate information fusion described above – combining multiple spaces with different characteristics – provides a setting for investigation of related issues. The recent survey [5] considers a wide range of examples and methodologies addressing this phenomenon in terms of *transfer learning*, *domain adaptation*, *multitask learning*, etc. The recent special issue [6] is devoted entirely to dimensionality reduction via subspace and submanifold learning. The majority of this article considers the Neyman-Pearson hypothesis testing setting, which provides clarity through the most straightforward of inference tasks. In Section 5.2 we briefly consider a *ranking* task.

Our dissimilarity-centric approach is motivated by the 2005 Pekalska and Duin book [7] on the dissimilarity representation for pattern recognition and the far-reaching success of multidimensional scaling methodologies [8, 9, 10, 11]

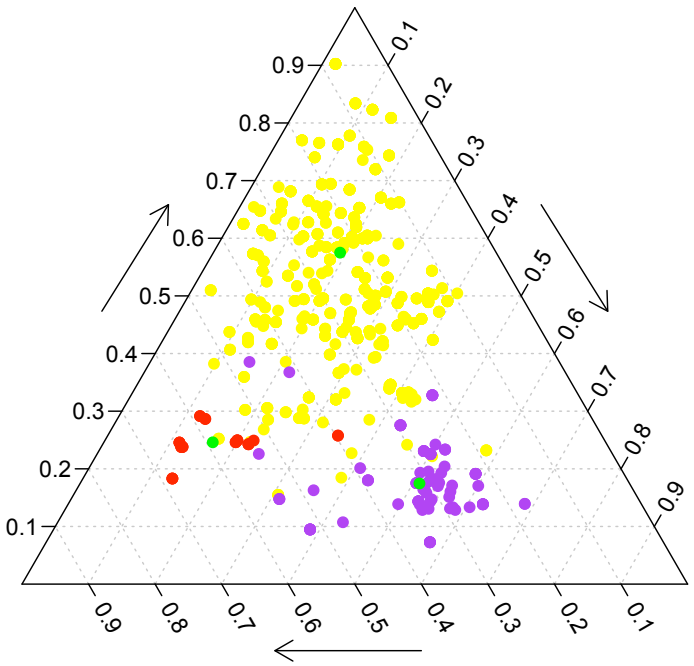

Figure 4: An example considering the FLD∘LSA projection into S^2 of multiple Enron email messages identified with three Linguistic Data Consortium (LDC) topics. The three colored scatterplots – yellow, red, purple – represent documents from the three topics; the green dots represent the topic means. We see that "matched", meaning "on the same topic", does mean something quite like $Dirichlet(r\boldsymbol{\gamma}_{topic} + \mathbf{1})$ in this case (but the variability "r" may be topic-dependent).

Combining information from disparate data sources when the information in the various spaces is fundamentally incommensurate – that is, a separate collection of useful features can be extracted from each space but their interpoint geometry precludes profitable alignment in a common space – is considered via Cartesian product space embedding in [12].

Preliminary development of our joint optimization methodology presented herein, as well as an application to *classification* tasks, is presented in [13].

1.7 Summary

In Section 2 we frame the problem as an optimization problem, and lay the groundwork for the methodologies proposed in Section 3. Section 4 illustrates the methodologies with instructive simulations that illustrate characteristic behavior; in particular, a simulation involving Dirichlet random variables sets the stage for the experimental examples on text documents presented in

Section 5. Finally, Section 6 provides discussion and suggestions for several areas of continuing research.

2 Fidelity and Commensurability

As suggested in Figure 1, our goal is to identify maps ρ_k taking Ξ_k to $\mathbb{R}^m$ (for some pre-specified m) such that (for $K = 2$) the power of the test, $P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A : \boldsymbol{y}_1 \nsim \boldsymbol{y}_2]$, is large, where the critical value c_α is determined by the null distribution of the test statistic and the allowable Type I error level α.

We proceed using ℓ_2 error for convenience and simplicity; clearly there is ample reason to consider other error criteria for particular applications. Similarly, we will assume symmetric dissimilarities δ_k.

The available matched points $\{\boldsymbol{x}_{ik}\}$ are used to identify appropriate maps ρ_k. Fidelity is how well the mapping $\boldsymbol{x}_{ik} \mapsto \widetilde{\boldsymbol{x}}_{ik}$ preserves original dissimilarities. The within-condition squared *fidelity error* is given by

$$\epsilon^2_{f_k} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (d(\widetilde{\boldsymbol{x}}_{ik}, \widetilde{\boldsymbol{x}}_{jk}) - \delta_k(\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk}))^2$$

for each k. If the fidelity error is large, then it is likely that the mapping does not capture aspects of original data that may be needed for inference.

On the other hand, even if all fidelity errors are small, inference may fail if $d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$ is large under the "matched" null hypothesis $H_0 : \boldsymbol{y}_1 \sim \boldsymbol{y}_2$. Commensurability is how well the mappings preserve matchedness; the between-condition squared *commensurability error* is given by

$$\epsilon^2_{c_{k_1 k_2}} = \frac{1}{n} \sum_{1 \le i \le n} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{ik_2}) - \delta_{k_1 k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{ik_2}))^2.$$

Alas, $\delta_{k_1 k_2}$ does not exist – we have no dissimilarity on $\Xi_{k_1} \times \Xi_{k_2}$. However, the concept of "matchedness" suggests that it might be reasonable to set $\delta_{k_1 k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{ik_2}) = 0$ for all i, k_1, k_2, in which case the commensurability error is the mean squared distance between matched points – the same criterion optimized by the Procrustes matching employed below.

There is also between-condition squared *separability error* given by

$$\epsilon^2_{s_{k_1 k_2}} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} (d(\widetilde{\boldsymbol{x}}_{ik_1}, \widetilde{\boldsymbol{x}}_{jk_2}) - \delta_{k_1 k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{jk_2}))^2.$$

However, it is less clear how to identify a reasonable stand-in for the $\delta_{k_1 k_2}$ terms in this expression. We will return to this issue when presenting our joint optimization inference methodology proposal in Section 3.3 below.

If all these errors are small – and if the target dimensionality is low enough so that estimation variance does not dominate (see e.g. [14] Section 3 and [15] Figure 12.1) – then successful inference in the target space may be achievable. The idea of the joint optimization method proposed in this manuscript (Section 3.3) is to attempt to minimize all three of these errors simultaneously.

3 Inference Methodologies

In this section we present three methodologies for performing our manifold matching inference – one which focuses on fidelity and is based on multidimensional scaling and Procrustes matching,

one which focuses on commensurability and is based on canonical correlation analysis, and then our proposal for joint optimization of fidelity and commensurability.

Before proceeding, we briefly review multidimensional scaling, Procrustes matching, and canonical correlation analysis.

Multidimensional scaling (MDS) takes an $n \times n$ dissimilarity matrix $\Delta = [\delta_{ij}]$ and produces a configuration of n points $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n$ in a target metric space endowed with distance function d such that the collection $\{d(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j)\}$ agrees as closely as possible with the original $\{\delta_{ij}\}$ under some specified error criterion; see for instance [8, 9, 10, 11]. For example, ℓ_2 (also known as "raw stress") MDS minimizes $\sum_{1 \le i < j \le n} (d(\widetilde{\boldsymbol{x}}_i, \widetilde{\boldsymbol{x}}_j) - \delta_{ij})^2$.

Out-of-sample embedding is used throughout this paper – given a configuration $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^n$ of the training observations and dissimilarities between test observations and the training observations, the test points are embedded into the existing configuration so as to be as ℓ_2-consistent as possible with these dissimilarities. This out-of-sample embedding can be one at a time, or jointly if the dissimilarities among multiple test observations are also available. Trosset and Priebe [16] present the out-of-sample methodology appropriate for classical MDS embeddings. We use raw stress embeddings herein, and the appropriate corresponding out-of-sample methodology is presented in [17].

Procrustes matching [18, 19, 20, 21] takes two matched collections $\widetilde{X}_1$ and $\widetilde{X}_2'$ of n points in $\mathbb{R}^m$ and finds the rigid motion transformation which optimally aligns the two collections. For example, ℓ_2 Procrustes minimizes the Frobenius norm $\|\widetilde{X}_1 - \widetilde{X}_2' Q\|_F$ over all $m \times m$ matrices Q such that $Q^T Q = I$. (We assume the dissimilarities have been scaled so that a scaling is not required in the Procrustes mapping. Thus Q defines a rigid motion mapping $\widetilde{X}_2'$ "onto" $\widetilde{X}_1$. We address this issue briefly in Section 6.)

Canonical correlation analysis (CCA) takes a collection X_1 of n_1 points in $\mathbb{R}^{m_1}$ and a collection X_2 of n_2 points in $\mathbb{R}^{m_2}$ and finds the pair of linear maps $U_1 : \mathbb{R}^{m_1} \to \mathbb{R}$ and $U_2 : \mathbb{R}^{m_2} \to \mathbb{R}$ which maximizes the correlation between $\widetilde{X}_1 = U_1(X_1)$ and $\widetilde{X}_2 = U_2(X_2)$. Performing m iterations of this procedure in the successive orthogonal subspaces yields a CCA procedure which maps to $\mathbb{R}^m$. See, for instance, [22, 23, 24].

Let us now consider these tools as building blocks for manifold matching inference.

3.1 Procrustes ∘ MDS

Multidimensional scaling yields low-dimensional embeddings. That is, $\Delta_1 \mapsto \widetilde{X}_1$ and $\Delta_2 \mapsto \widetilde{X}_2'$ yields $n \times m$ configurations. Procrustes$(\widetilde{X}_1, \widetilde{X}_2')$ yields

$$Q^* = \arg \min_{Q^T Q = I} \|\widetilde{X}_1 - \widetilde{X}_2' Q\|_F.$$

Given $\delta_k(\boldsymbol{y}_k, \boldsymbol{x}_{ik})$, $i = 1, \ldots, n$ for $k = 1, 2$, out-of-sample embedding of the test data gives $\boldsymbol{y}_1 \mapsto \widetilde{\boldsymbol{y}}_1$, $\boldsymbol{y}_2 \mapsto \widetilde{\boldsymbol{y}}_2'$ where the embedded points are chosen so that their distances to $\widetilde{\boldsymbol{x}}_{ik}$ agree as closely as possible with the available dissimilarities. Using the rigid motion transformation obtained in the Procrustes step, both $\widetilde{\boldsymbol{y}}_1$ and $\widetilde{\boldsymbol{y}}_2 = ((\widetilde{\boldsymbol{y}}_2')^T Q^*)^T$ are in $\mathbb{R}^m$ with same coordinate system. Thus inference may proceed by rejecting for large values of $d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$. We dub this separate embedding approach "Procrustes composed with multidimensional scaling", or "$p \circ m$".

From an inspection of the raw stress multidimensional scaling criterion function, it follows immediately that the $\Delta_k \mapsto \widetilde{X}_k$ mappings minimize fidelity error. Thus we have established the following result:

Theorem 1: $p \circ m$ optimizes fidelity *without regard for commensurability.*

That is, the maps ρ_k are identified separately, with no concern for whether the commensurability optimization in the Procrustes step will be able to provide a good alignment.

3.2 Canonical Correlation

Since canonical correlation begins with Euclidean data, the first step of this methodology necessarily involves multidimensional scaling. This appears similar to Procrustes $\circ$ MDS above, but in this case no attempt is made to achieve meaningful dimensionality reduction. Multidimensional scaling yields high-dimensional embeddings, $\Delta_1 \mapsto X_1'$ and $\Delta_2 \mapsto X_2'$, but in this case these maps are to the highest-dimensional space possible, $\mathbb{R}^{n-1}$ in general. Canonical correlation finds linear maps to $\mathbb{R}^m$, $U_1 : X_1' \mapsto \widetilde{X}_1$ and $U_2 : X_2' \mapsto \widetilde{X}_2$, to maximize correlation. Again, out-of-sample embedding yields $(n-1)$-dimensional points $\boldsymbol{y}_1 \mapsto \boldsymbol{y}_1'$, $\boldsymbol{y}_2 \mapsto \boldsymbol{y}_2'$. Then $\widetilde{\boldsymbol{y}}_1 = U_1^T \boldsymbol{y}_1'$ and $\widetilde{\boldsymbol{y}}_2 = U_2^T \boldsymbol{y}_2'$ can be directly compared. An investigation of the correlation criterion function shows that the CCA maps U_1 and U_2 minimize commensurability error, subject to linearity. Thus there is no need for Procrustes in this case, and once again inference may proceed: reject for large values of $d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$. We dub this approach "*cca*".

From the equivalence of the correlation objective function and commensurability error, we have established the following result:

Theorem 2: *cca* optimizes commensurability *without regard for fidelity.*

That is, the maps ρ_k are identified jointly, but with no concern for fidelity of the individual embeddings (beyond linearity).

3.3 Omnibus Embedding

In response to the optimization objectives of the two methodologies presented above – one considering fidelity only and the other considering commensurability only – we develop an omnibus embedding methodology explicitly focused on the joint optimization of fidelity and commensurability.

$$\overset{2n\times 2n}{M} = \begin{bmatrix} \overset{n\times n}{\Delta_1} & \overset{n\times n}{W} \\ W^T & \overset{n\times n}{\Delta_2} \end{bmatrix} \begin{matrix} \overset{n\times 1}{u_1} \; \overset{n\times 1}{u_2} \\ \\ \overset{n\times 1}{v_1} \; \overset{n\times 1}{v_2} \end{matrix}$$

$$\begin{matrix} y_1 \\ y_2 \end{matrix} \quad \begin{matrix} u_1^T \\ u_2^T \end{matrix} \quad \begin{matrix} v_1^T \\ v_2^T \end{matrix}$$

Figure 5: Depiction of the $2n \times 2n$ omnibus dissimilarity matrix M, including imputed dissimilarities $W = [\delta_{12}(\boldsymbol{x}_{i1}, \boldsymbol{x}_{j2})]$ and out-of-sample test data $\boldsymbol{y}_1, \boldsymbol{y}_2$.

Under the "matched" assumption, we *impute* dissimilarities $W = [\delta_{12}(\boldsymbol{x}_{i1}, \boldsymbol{x}_{j2})]$ to obtain a $2n \times 2n$ *omnibus dissimilarity matrix* M. See Figure 5, which depicts M as a block matrix consisting of the $n \times n$ dissimilarities matrices Δ_1 and Δ_2 on the diagonal and W as the $n \times n$ off-diagonal block. (This generalizes immediately to $K > 2$.) As discussed above, it seems reasonable under H_0 to set the diagonal elements $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{ik_2})$ of W to zero. (Notice, however, that $\delta_{k_1k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{ik_2}) = 0$ for $k_1 \neq k_2$ is not necessarily "truth." For instance, the Dirichlet setting of Section 1.5 with $r < \infty$ would have non-zero elements for $diag(W)$. Still, this "shrinkage" of $diag(W)$ to zero seems reasonable.) As for the off-diagonal elements of W, we argue that either leaving them as missing data unused in the subsequent optimization or letting $W = (\Delta_1 + \Delta_2)/2$ are reasonable suggestions; we will return to this imputation issue later. Once we have settled on W, our approach considers MDS embedding of M as $2n$ points in $\mathbb{R}^m$ – zeros on the diagonal of W act to force matched points to be embedded near each other. It is clear that raw stress MDS applied to M has as its objective function precisely $\epsilon_{f_1}^2 + \epsilon_{f_2}^2 + \epsilon_{c_{12}}^2 + \epsilon_{s_{12}}^2$. If $diag(W) = 0$ and the off-diagonal elements are treated as missing and ignored in the optimization, then this objective function reduces to a consideration of just fidelity and commensurability.

Let $u_{i1} = \delta_1(\boldsymbol{y}_1, \boldsymbol{x}_{i1})$ and $v_{i2} = \delta_2(\boldsymbol{y}_2, \boldsymbol{x}_{i2})$. Under H_0, impute $v_{i1} = \delta_{12}(\boldsymbol{y}_1, \boldsymbol{x}_{i2})$ and $u_{i2} = \delta_{12}(\boldsymbol{y}_2, \boldsymbol{x}_{i1})$ via $\boldsymbol{v}_1 = \boldsymbol{u}_2 = (\boldsymbol{u}_1 + \boldsymbol{v}_2)/2$. Out-of-sample embedding of $(\boldsymbol{u}_1^T, \boldsymbol{v}_1^T)^T$ and $(\boldsymbol{u}_2^T, \boldsymbol{v}_2^T)^T$ yields $\widetilde{\boldsymbol{y}}_1$ and $\widetilde{\boldsymbol{y}}_2$. Reject for large values of $d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$. We dub this omnibus embedding approach for joint optimization of fidelity and commensurability "*jofc*".

Obviously, the choice of W is key for this joint optimization. Also, note that weights can be incorporated into the MDS optimization criterion; this weighting can become quite elaborate, but in its simplest form it yields a more general tradeoff between fidelity and commensurability via $\omega(\epsilon_{f_1}^2 + \epsilon_{f_2}^2) + (1 - \omega)\epsilon_{c_{12}}^2$.

4 Illustrative Simulation

In this section we present an illustrative Dirichlet simulation which helps to elucidate when and why our joint optimization methodology is superior to either version of separate optimization.

4.1 Dirichlet Product Model

We describe a probability model with parameters p, q, r, a, and K.

Let $\Xi_k = S^{p+q}$, $k = 1, 2$. Here the simplex S^p encodes "signal" and the simplex S^q encodes "noise". That is, on S^p we let $\boldsymbol{\gamma}_i \overset{iid}{\sim} Dirichlet(\mathbf{1})$ and mutually independent $X_{ik}^1 \sim Dirichlet(r\boldsymbol{\gamma}_i + \mathbf{1})$ (signal, as in Section 1.5) while on S^q we let $X_{ik}^2 \overset{iid}{\sim} Dirichlet(\mathbf{1})$ (pure noise). For $a \in [0, 1]$, let $X_{ik} = [(1 - a)X_{ik}^1, aX_{ik}^2]$ – the concatenation of (weighted) signal and noise dimensions. The resultant distribution for $(X_{i1}, \cdots, X_{iK})$ is denoted by $F_{p,q,r,a,K}$, and $F_{p,q,r,a,K|\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n}$ denotes the distribution conditional on the location of the $\boldsymbol{\gamma}_i$.

4.2 Testing

For each of n_{mc} Monte Carlo replicates ($n_{mc} = 1000$ in the simulations), we generate n matched pairs according to the Dirichlet product model distribution $F_{p,q,r,a,K=2}$ by first generating $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$ and then, conditional on the collection $\{\boldsymbol{\gamma}_i\}$, generating the matched pair (X_{i1}, X_{i2}). Embeddings are defined for each of the three competing methodologies based on this matched training data. For each test datum under H_0, one new $\boldsymbol{\gamma}$ is generated, a matched pair is generated, out-of-sample embedding is performed, and the statistic $T = d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$ is calculated; this is repeated s times independently ($s = 1000$ in the simulations) and the critical value c_α for the allowable

Type I error level α is determined based on the Monte Carlo estimate of null distribution of T. Then *un*matched pairs are generated, out-of-sample embedding is performed, and the statistic T is calculated for test data under H_A; this provides an estimate of the conditional power $P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n]$.

We perform n_{mc} Monte Carlo replicates to integrate out the $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n$, yielding comparative power estimates. We also investigate conditional power for particular collections $\{\boldsymbol{\gamma}_i\}$, in order to better understand precisely when and why our joint optimization methodology is superior to either version of separate optimization.

4.3 Results

Figure 6 presents results from our Dirichlet product model. $K = 2$, with $p = 3, q = 3, r = 100, a = 0.1$. The target dimension is $m = 2$. We use $n = 100$. The allowable Type I error level α is plotted against power $\beta = P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A]$. The results are based on $n_{mc} = 1000$ Monte Carlo replicates with $s = 1000$; the differences in the curves are statistically significant. In this case, *jofc* with $W = (\Delta_1 + \Delta_2)/2$ is superior to both $p \circ m$ and *cca*.

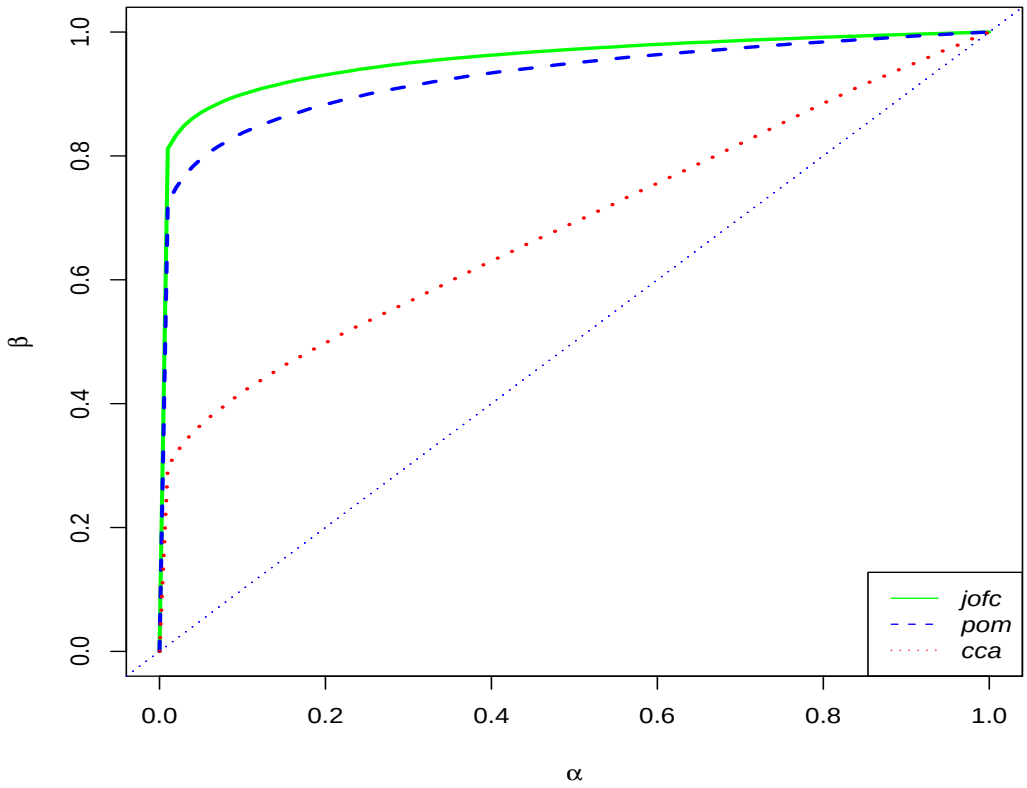

Figure 6: Dirichlet product model simulation results plotting the Type I error level α against power $\beta = P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A]$, indicating that *jofc* is superior to both $p \circ m$ and *cca*. See text for description.

4.4 Analysis

The Dirichlet product model is designed specifically to illustrate when and why *jofc* is superior to both *p∘m* and *cca* in terms of fidelity and commensurability.

If q is large with respect to the target dimensionality m, then with high probability *cca* will identify a $m-$dimensional subspace in the "noise" simplex S^q with spurious correlation. This phenomenon requires only that $a > 0$. In this event, the out-of-sample embedding will produce arbitrary $\widetilde{\boldsymbol{y}}_1$ and $\widetilde{\boldsymbol{y}}_2$, even under H_0. Thus the null distribution of the test statistic will be inflated by these spurious correlations. If the allowable Type I error level is smaller than the probability of inflation, then the power of the *cca* method will be negatively affected.

If a is small and $m \leq p$, then with high probability the $m-$dimensional subspaces identified by the MDS step will come from the "signal" simplex S^p. If $m < p$, then with positive probability, these two subspaces, identified separately in *p∘m*, will be geometrically incommensurate (see Figure 7). Thus the null distribution of the test statistic will be inflated by these incommensurate cases. If the allowable Type I error level α is smaller than the probability of inflation, then the power of the *p∘m* method will be negatively affected.

For large q and small a, the two phenomena described above occur in the same model. The *jofc* method is not susceptible to either phenomenon: incorporating fidelity into the objective function obviates the spurious correlation phenomenon, and incorporating commensurability into the objective function obviates the geometric incommensurability phenomenon. Thus we can establish that, for a range of Dirichlet product model distributions, *jofc* is superior to both *p∘m* and *cca*.

Theorem 3: Let $m \in \{1, \cdots, \min\{p-1, q\}\}$, $a \in (0, 1/2)$, and $r \in (0, \infty)$. Then for large q, small a, and large r, there exists allowable Type I error level $\alpha > 0$ such that the Dirichlet product model distribution $F_{p,q,r,a,K=2}$ with target dimensionality m yields power $\beta_{jofc} > \max\{\beta_{p\circ m}, \beta_{cca}\}$, where power $\beta = P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A]$ for the various testing methodologies *jofc*, *p∘m*, and *cca*.

Proof: Let b_1 denote the probability that *cca* suffers from the spurious correlation phenomenon, and let b_2 denote the probability that *p∘m* suffers from the geometric incommensurability phenomenon. Then $q \gg p$ implies that *cca* suffers from the spurious correlation phenomenon with high probability and thus $b_1 \approx 1$ and $\beta_{cca} \approx \alpha$. For $a \approx 0$ and r sufficiently large, *jofc* and *p∘m* identify approximately the same embeddings *except* for the cases in which *p∘m* suffers from the incommensurability phenomenon. Thus the null distribution of $T = d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$ for *jofc* is approximately point mass at zero while the null distribution of T for *p∘m* has b_2 mass $\gg 0$. Hence $\alpha \approx b_2/2$ yields $\beta_{jofc} \approx 1$ while $\beta_{p\circ m} \approx 1/2$.■

Delving into our simulation results via investigation of conditional power $P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n]$, it is apparent that the superiority of *jofc* is indeed due to occurrences of the phenomena described above – individual Monte Carlo replicates (particular selections of the $\{\boldsymbol{\gamma}_i\}$, essentially) are identified in which the spurious correlation phenomenon causes poor performance for *cca* or the incommensurability phenomenon causes poor performance for *p∘m* but in which *jofc* is unaffected.

We note that the Dirichlet product model introduced here as an aid in understanding when and why *jofc* is superior to both *p∘m* and *cca* does in fact (loosely) model general high-dimensional real data scenarios: many dimensions consisting mostly of noise along with a few signal dimensions.

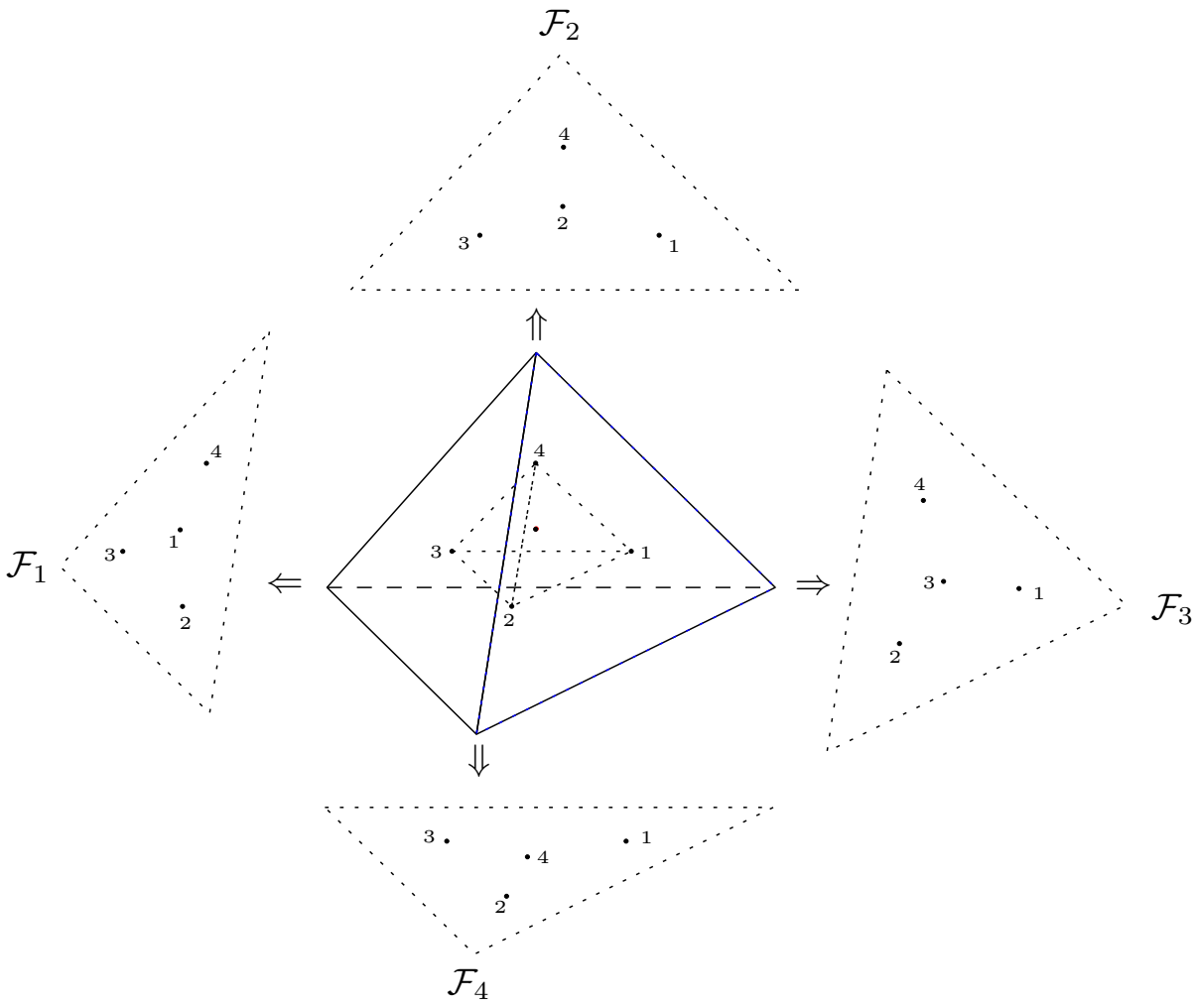

Figure 7: Idealization of the incommensurability phenomenon: for a symmetric collection $\{\gamma_1, \gamma_2, \gamma_3, \gamma_4\}$ in the simplex S^3, all four of the facet projections have the same fidelity and are geometrically incommensurable with one another.

4.5 Gaussian Model

A Gaussian model, analogous to the Dirichlet product model investigated above, is constructed here to provide a sense of the generality of models with many dimensions consisting mostly of noise along with a few signal dimensions.

We consider p-dimensional means $\boldsymbol{\mu}_i \overset{iid}{\sim} \mathcal{N}\left(\vec{0}, I_p\right)$, $i = 1, \cdots, n$, analogous to the $\boldsymbol{\gamma}_i$ from the Dirichlet model. Matchedness arises from independent $X_{ik}^1 \sim \mathcal{N}\left(\boldsymbol{\mu}_i, r^{-1} I_p\right)$, $i = 1, \ldots, n$, $k = 1, \ldots K$, for $r \in (0, \infty)$; as r increases, the degree of matchedness increases. As before, we have q-dimensional "noise" vectors $X_{ik}^2 \overset{iid}{\sim} \mathcal{N}\left(\vec{0}, I_q\right)$. Again, for $a \in [0, 1]$, $X_{ik} = [(1-a)X_{ik}^1, aX_{ik}^2]$ represents the concatenation of (weighted) signal and noise dimensions. As with the Dirichlet product model, both the spurious correlation phenomenon and the geometric incommensurability phenomenon are present in this Gaussian model.

Figure 8 presents simulation results for this Gaussian model, entirely analogous to those depicted in Figure 6.

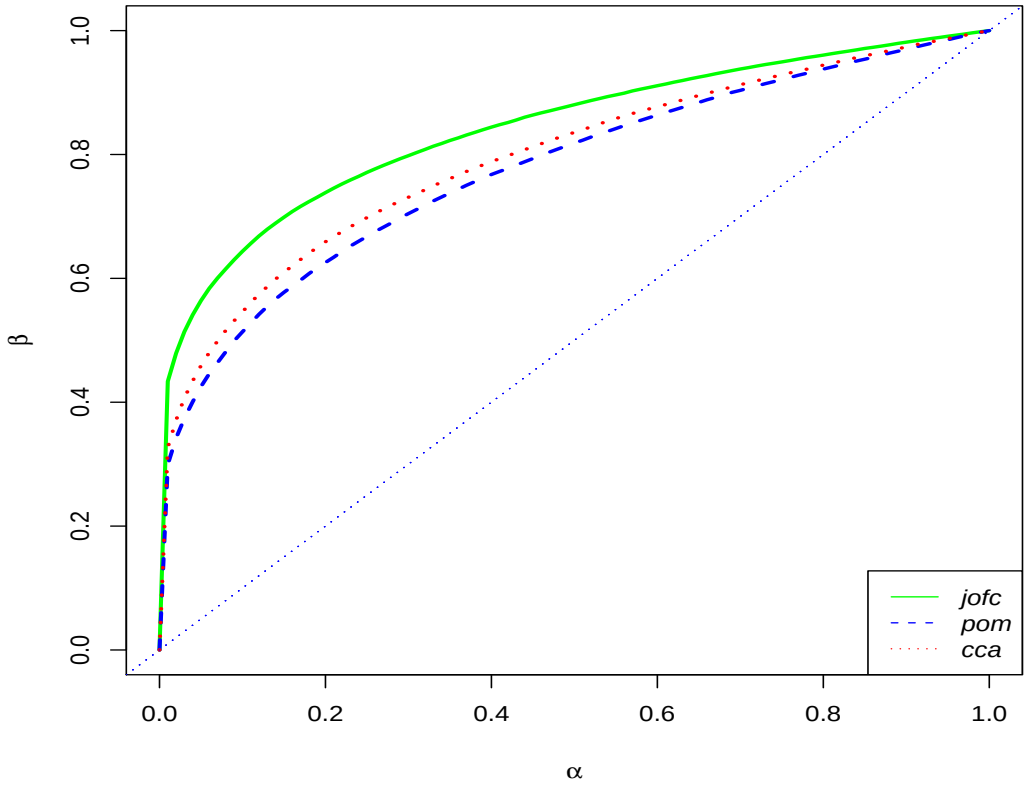

Figure 8: Gaussian model simulation results plotting the Type I error level α against power $\beta = P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A]$, indicating *jofc* is superior to both $p \circ m$ and *cca*, entirely analogous to those presented for the Dirichlet product model in Figure 6.

5 Experimental Results

5.1 Testing

A collection of documents $\{\boldsymbol{x}_{i1}\}_{i=1}^{n}$ are collected from the English Wikipedia, corresponding to the directed 2-neighborhood of the document "Algebraic Geometry." This yields $n = 1382$ and, through Wikipedia's own 1-1 correspondence, the associated French documents $\{\boldsymbol{x}_{i2}\}_{i=1}^{n}$. For dissimilarity matrices Δ_k, $k = 1, 2$, we use the Lin & Pantel discounted mutual information [25, 26] and cosine dissimilarity $\delta_k(\boldsymbol{x}_{ik}, \boldsymbol{x}_{jk}) = 1 - (\boldsymbol{x}_{ik} \cdot \boldsymbol{x}_{jk})/(\|\boldsymbol{x}_{ik}\|_2 \|\boldsymbol{x}_{jk}\|_2)$.

Our results are obtained by repeatedly randomly holding out four documents – two matched pairs – and identifying the embeddings via *cca*, *p*$\circ$*m*, and *jofc* based on the remaining $n = 1380$ matched pairs. The two sets of held-out matched pairs are used as $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, via out-of-sample embedding, to estimate the null distribution of the test statistic $T = d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2)$. This allows us to estimate critical values for any specified Type I error level. Then the two sets of held-out *un*matched pairs are used as $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, via out-of-sample embedding, to estimate power. Target dimensionality m is determined by the Zhu and Ghodsi automatic dimensionality selection method [27], resulting in $m = 6$ for this data set.

Figure 9 plots the allowable Type I error level against power. These experimental results indicate that *jofc* is superior to both *p*$\circ$*m* and *cca*, and are entirely analogous to the simulation results presented above.

5.2 Ranking

Here we consider a *ranking* task in which matched training data exists in disparate spaces Ξ_1 and Ξ_2, but test observation $\boldsymbol{y}_2$ will be observed in space Ξ_2. The task is to find the match for $\boldsymbol{y}_2$ amongst a candidate collection $\mathcal{C} = \{\boldsymbol{y}_{11}, \cdots, \boldsymbol{y}_{z1}\} \subset \Xi_1$ of $z > 1$ possibilities. Using the training set of matched observations, we identify the embeddings via *cca*, *p*$\circ$*m*, and *jofc*, and out-of-sample embedding then yields $\widetilde{\boldsymbol{y}}_2$ and $\widetilde{\mathcal{C}} = \{\widetilde{\boldsymbol{y}}_{11}, \cdots, \widetilde{\boldsymbol{y}}_{z1}\}$. The *rank* r^* of the one true match to $\boldsymbol{y}_2$ amongst the candidate collection $\mathcal{C}$ in terms of $\{d(\widetilde{\boldsymbol{y}}_{\zeta 1}, \widetilde{\boldsymbol{y}}_2)\}_{\zeta=1}^{z}$ is our measure of performance; $r^* = 1$ represents perfect performance, $r^* = z/2$ represents chance, and $r^* = z$ is the worst possible.

For this experiment we consider a different collection of Wikipedia documents: all English/Persian (Farsi) matched pairs (matched, again, through Wikipedia's own 1-1 correspondence) for which both documents in the pair contain at least 500 total words and at least 100 distinct words. There are 2448 such pairs. (The word-count restrictions are to ensure that the documents are legitimate articles, rather than "stubs" – place-holders for future articles on the topic.)

Figures 10 and 11 present notched boxplot experimental results wherein we repeatedly hold out $z = 1000$ matched pairs from the training set. (Recall that non-overlapping notches implies a statistically significant difference of means.) Figure 10 depicts r^* as a function of target dimension m for *jofc* (gray) and *p*$\circ$*m* (white). Performance improves for both methods as m increases from 5 to 25, with *jofc* superior. Performance levels off after $m = 30$ (and degrades significantly for $m > 50$). Figure 11 depicts difference in ranks, $r^*_{p\circ m} - r^*_{jofc}$; differences greater than 0 indicate *jofc* superiority.

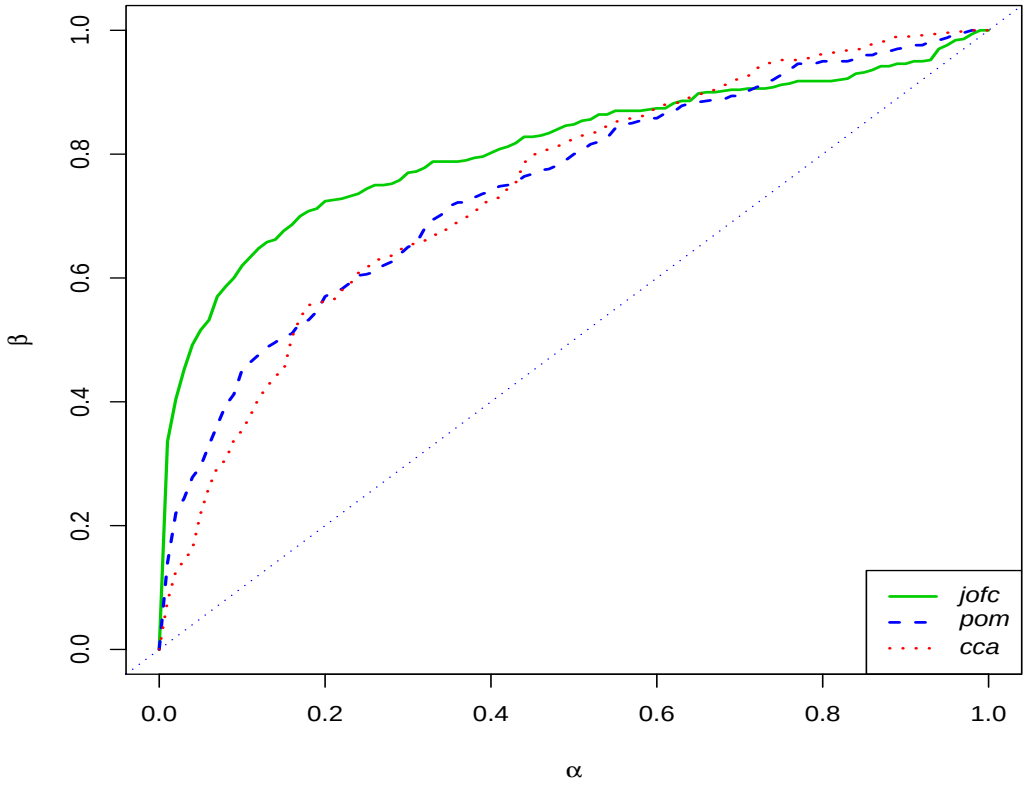

Figure 9: Experimental results on English/French Wikipedia documents plotting the Type I error level α against power $\beta = P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A]$, indicating *jofc* is superior to both $p \circ m$ and *cca*. See text for description.

6 Discussion and Conclusions

We have presented a complete methodological core for manifold matching via joint optimization of fidelity and commensurability and comprehensive comparisons with either version of separate optimization. Continuing research includes comparison with other standard competing methodologies, variations and generalizations of our omnibus embedding methodology, and further theoretical developments.

Here we discuss a few of the most pressing issues.

$K > 2$ Conditions

It is straightforward to generalize the omnibus dissimilarity matrix M to the case of $K > 2$ conditions.

Pre-Scaling the Δ_k

The *scale* of the various dissimilarities has been assumed to be consistent. For Dirichlet data, this assumption is warranted; however, pre-scaling of the Δ_k prior to constructing M is imperative

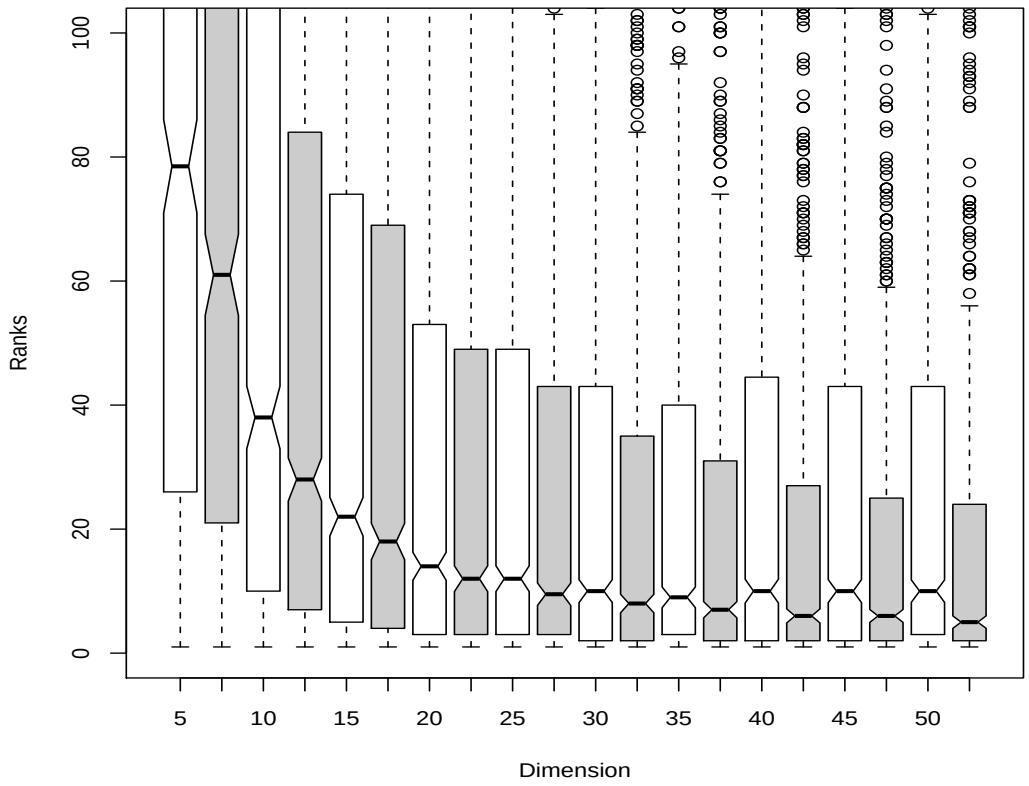

Figure 10: Comparative rank experimental results depicting the rank r^* of the one true match to test observation $\boldsymbol{y}_2$ amongst the candidate collection $\mathcal{C}$ in terms of $\{d(\widetilde{\boldsymbol{y}}_{\varsigma 1}, \widetilde{\boldsymbol{y}}_2)\}_{\varsigma=1}^{z}$ as a function of target dimension m. For each $m \in \{5, 10, 15, \cdots, 50\}$, there are two boxplots. These results indicate that *jofc* (gray) is superior to $p \circ m$ (white) on this data set. With $z = 1000$, both methods perform much better than chance ($r^* = z/2$), although performance does not achieve perfection ($r^* = 1$). See text for description.

for the general case.

MDS Objective

Our omnibus embedding methodology can be employed with MDS criteria other than raw stress; the ℓ_2 criterion provides direct correspondence to fidelity and commensurability. Weighted ℓ_2 is straightforward. Other MDS minimization objectives have been studied in depth, and should in particular circumstances provide superior performance.

Imputation of W

It seems reasonable under H_0 to set the diagonal elements $\delta_{k_1 k_2}(\boldsymbol{x}_{ik_1}, \boldsymbol{x}_{ik_2})$ of W to zero. Recall, however, that this is not necessarily "truth;" the Dirichlet setting of Section 1.5 with $r < \infty$ would have non-zero elements for $diag(W)$. Still, this shrinkage of $diag(W)$ to zero seems reasonable.

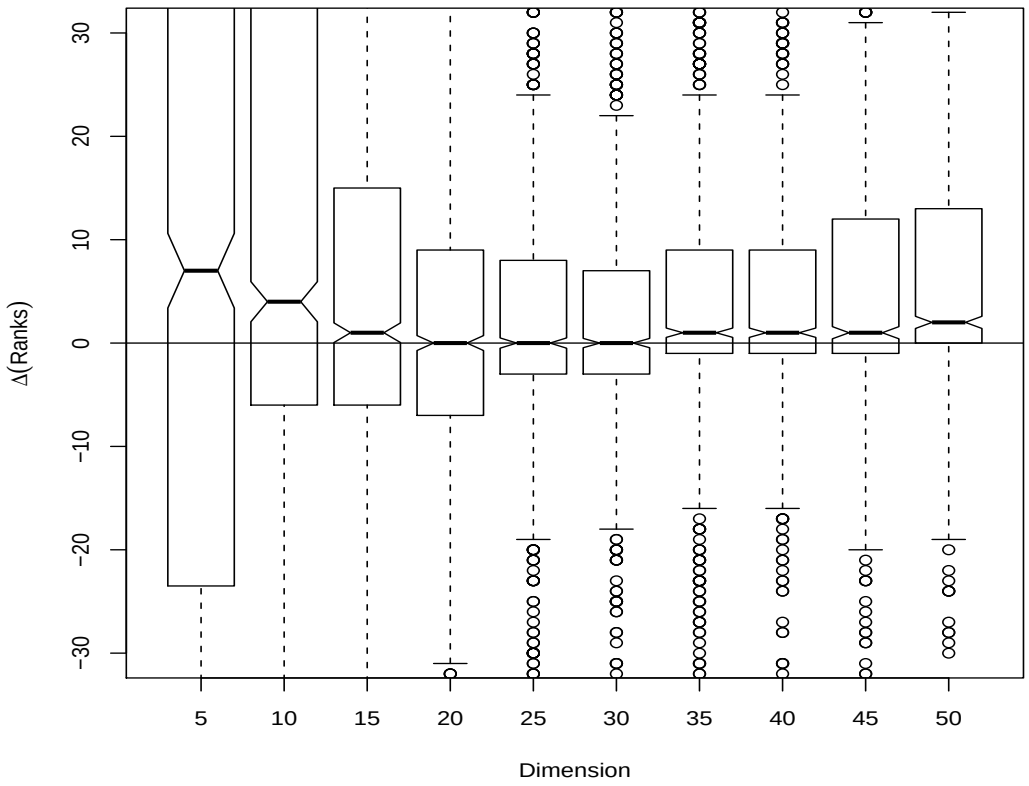

Figure 11: Comparative rank experimental results depicting difference in ranks $r^*_{p\circ m} - r^*_{jofc}$; differences greater than 0 indicate *jofc* superiority. See text for description.

However, there may be cases for which imputing non-zero values would be appropriate; for example, if information is available suggesting that some matchings are unreliable, then it might be advantageous to use larger values for these matchings.

As for the off-diagonal elements of W, we have argued that either leaving them as missing data unused in the subsequent optimization or letting $W = (\Delta_1 + \Delta_2)/2$ are reasonable suggestions. We believe that more elaborate imputation should provide superior performance. In particular, it seems clear that choosing $\lambda \in [0, 1]$ and setting $W = \lambda\Delta_1 + (1-\lambda)\Delta_2$ or $W = (\lambda\Delta_1^2 + (1-\lambda)\Delta_2^2)^{1/2}$ will be preferable in certain circumstances.

Model Selection: The Choice of Target Dimensionality m

We have assumed throughout that $\mathcal{X} = \mathbb{R}^m$ for some pre-specified target dimension m. First, we note that, in general, embedding into target spaces other than Euclidean is possible and sometimes productive. More pressing is the necessity, in many applications, for data-driven choice of target dimension. This is in general a vexing model selection task – the bias-variance trade-off. Of course, $m = 1$ generally induces significant model bias and $m = n - 1$ generally admits excessive estimation variance, as characterized in [15] Figure 12.1. Many dimensionality selection methods based on the principle of diminishing returns in terms of variance explained

are available – in Section 5.1 we made use of the method proposed in [27], and in 5.2 we presented results as a function of m. A dimensionality selection methodology specifically designed for use with our omnibus embedding methodology is of significant interest.

One illustrative point in this regard is that the general commensurate-space approach considered throughout this article – for all three approaches *jofc*, *p∘m*, and *cca* – adds a further complication with respect to identification of optimal target dimension: the optimal target dimension m_k^* for the various Δ_k will not the be same. This adds to the degree of difficulty in designing methods for identifying the optimal common-space target dimension m^*.

Learning the π_k

We have assumed that the maps π_k from object space Ξ to the conditional spaces Ξ_k are fixed (see Figure 1). Indeed, Ξ and the π_k have been treated as notional only. In some circumstances, it may be possible to use performance analyses to glean information concerning the induced conditional distributions and profitably adjust the π_k, in a manner analogous to fusion frames [28].

Fast Omnibus Embedding

Out-of-sample embedding of test data precludes re-learning the mappings for each inference. More importantly, it is straightforward to make a version of our omnibus embedding methodology fast ($O(n)$). Making an *effective* fast version requires numerous methodological choices for various stages of *jofc*.

Commensurability Error vs Hausdorff Distance on $G_{p,m}$

In the simple setting of Euclidean spaces Ξ_k, the *p∘m* methodology yields two elements of the Grassmann space $G_{p,m}$ of m-dimensional subspaces of $\mathbb{R}^p$. This space is a manifold under the Hausdorff distance $2\sin(\theta/2)$, where θ is the canonical angle between subspaces [29]. Under special conditions the Hausdorff distance between *p∘m*'s two subspaces and the commensurability error between their respective embeddings are closely related.

See Figure 12 for a first example, from the Dirichlet product model simulation presented in Figure 6. Each point in Figure 12 represents a Monte Carlo replicate. We note that the Hausdorff distance between *p∘m*'s two subspaces and the commensurability error between their respective embeddings are strongly correlated. Furthermore, the red points represent replicates for which the conditional power $P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n]$ is low – predominantly those replicates for which Hausdorff distance and commensurability error are large. This demonstrates the effect of the incommensurability phenomenon on *p∘m*. The *jofc* embeddings are not subject to this deleterious phenomenon.

Additional investigations concerning the superiority of *jofc* to *p∘m* due to the incommensurability phenomenon involve this relationship between Hausdorff distance and commensurability error. Significantly more involved investigations are required when, as is the case for proper text document analysis, one uses a more appropriate dissimilarity (Hellinger distance, or more generally α-divergence) on the simplex.

Three-Way MDS

Three-way MDS (see, for instance, [11]) addresses a problem superficially similar to joint optimization of fidelity and commensurability, in which a single configuration and two transformation matrices are identified from two dissimilarity matrices Δ_1, Δ_2. It may be of interest to compare

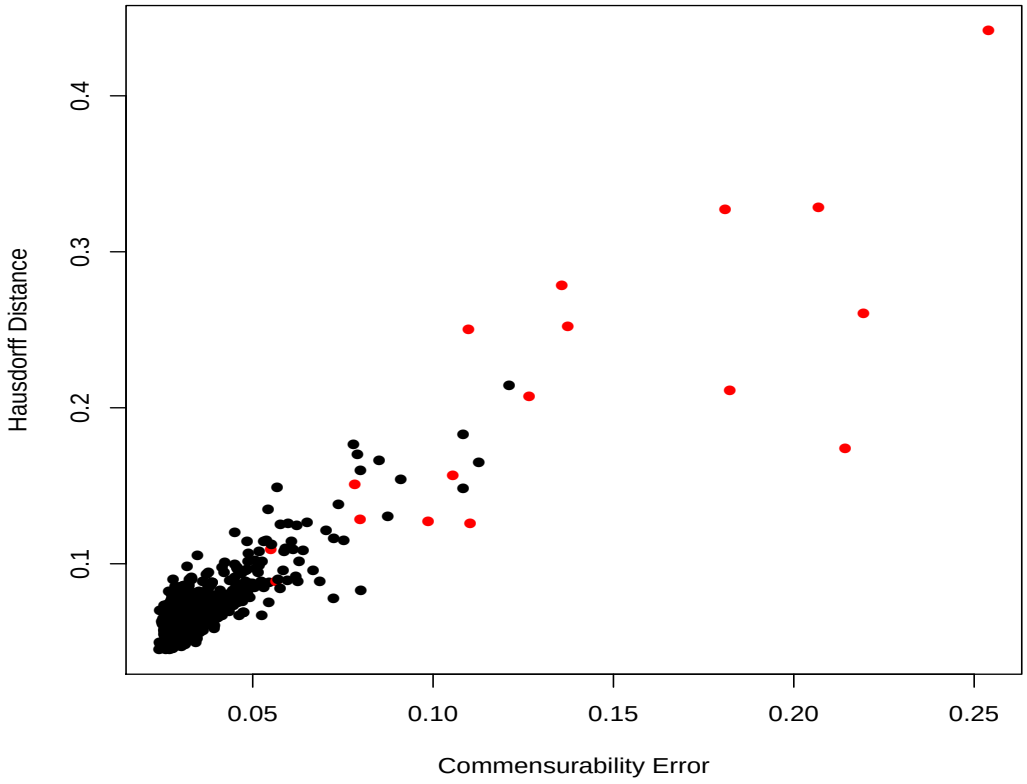

Figure 12: Commensurability error and Hausdorff distance on the Grassmannian Manifold for our Dirichlet product model simulation (Figure 6). Strong correlation is evident. Furthermore, the red points represent replicates for which the conditional power $P[d(\widetilde{\boldsymbol{y}}_1, \widetilde{\boldsymbol{y}}_2) > c_\alpha | H_A, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_n]$ is low – predominantly those replicates for which Hausdorff distance and commensurability error are large.

and contrast our omnibus embedding methodology with various instantiations of three-way MDS – particularly the identity model presented in [30].

6.1 Conclusions

In conclusion, we have presented an omnibus embedding methodology for joint optimization of fidelity and commensurability that allows us to address the manifold matching problem by jointly identifying embeddings of multiple spaces into a common space. Such a joint embedding facilitates statistical inference in a wide array of disparate information fusion applications. We have investigated this methodology in the context of simple statistical inference tasks, and compared and contrasted with competing fidelity-only and commensurability-only methodologies, demonstrating the superiority of our joint optimization.

We have focused on a simple setting and simple choices for various methodological options. Many variations and generalizations are possible, but the presentation here provides the core methodological instantiation.

References

[1] M. W. Berry. *Survey of Text Mining I: Clustering, Classification, and Retrieval (No. 1).* Springer, 2003.

[2] M. W. Berry. *Survey of Text Mining II: Clustering, Classification, and Retrieval (No. 2).* Springer, 2007.

[3] M. W. Berry and J. Kogan. *Text Mining: Applications and Theory.* Wiley, 2010.

[4] D. J. Hand. Classifier technology and the illusion of progress. *Statistical Science* pp. 1–34, 2006.

[5] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359, 2010.

[6] Y. Ma, P. Niyogi, G. Sapiro, and R. Vidal. Dimensionality reduction via subspace and submanifold learning. *IEEE Signal Processing Magazine* 28(2):14–126, 2011.

[7] E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications (Machine Perception and Artificial Intelligence).* World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005.

[8] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 1952.

[9] W. Torgerson. *Theory and Methods of Scaling.* John Wiley & Sons, 1958.

[10] T. Cox and M. Cox. *Multidimensional Scaling.* Chapman and Hall, 2001.

[11] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications.* Springer-Verlag, 2005.

[12] Z. Ma, A. Cardinal-Stakenas, Y. Park, M. W. Trosset, and C. E. Priebe. Dimensionality reduction on the cartesian product of embeddings of multiple dissimilarity matrices. *Journal of Classification* 27(3):307–321, 2010.

[13] Z. Ma, D. J. Marchette, and C. E. Priebe. Fusion and inference from multiple data sources in commensurate space. *Statistical Analysis and Data Mining*, accepted for publication, 2011.

[14] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(1):4–37, 2000.

[15] L. Devroy, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[16] M. W. Trosset and C. E. Priebe. The out-of-sample problem for classical multidimensional scaling. *Computational Statistics and Data Analysis* 52(10):4635–4642, 2008.

[17] Z. Ma and C. E. Priebe. Out-of-sample embedding using iterative majorization. submitted for publication, 2010.

[18] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science* 4(2):87–99, 1989.

[19] F. Bookstein. *Morphometric tools for landmark data*. Cambridge University Press, 1991.

[20] K. Mardia and I.L.Dryden. *Statistical Shape Analysis*. Wiley, Chichester, 1998.

[21] J. Gower and G. Dijksterhuis. *Procrustes Problems*. Oxford University Press, 2004.

[22] H. Hotelling. Relations between two sets of variates. *Biometrika* 28(3-4):321, 1936.

[23] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press, 1980.

[24] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation* 16(12):2639, 2004.

[25] D. Lin and P. Pantel. Concept discovery from text. *Proceedings of the 19th International Conference on Computational Linguistics* pp. 1–7, 2002.

[26] P. Pantel and D. Lin. Discovering word senses from text. *Proceedings of the ACM SIGKDD Conference on Knowldedge Discovery and Data Mining* pp. 613–619, 2002.

[27] M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis* 51:918–930, 2006.

[28] R. Calderbank, P. G. Casazza, A. Heinecke, G. Kutyniok, and A. Pezeshki. Sparse fusion frames: Existence and construction. *Adv. Comput. Math.*, to appear.

[29] L. Qiu, Y. Zhang, and C.-K. Li. Unitarily invariant metrics on the grassmann space. *SIAM J. Matrix Anal. Appl.* 27(2):507–531, 2005.

[30] J. Commandeur and W. Heiser. Mathematical derivations in the proximity scaling (proxscal) of symmetric data matrices. *Research Report RR-93-04, Department of Data Theory, Leiden University*, 1993.